DOCUMENT RESUME

ED 405 880 IR 056 309

AUTHOR Armstrong, C. J.
TITLE Database Quality: Label or Liable.
PUB DATE [95]
NOTE 6p.; Paper presented at the Northumbria International Conference on Performance Measurement in Libraries and Information Services (1st, Northumberland, England, August 30-September 4, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Accrediting Agencies; Clearinghouses; Coding; Consumer Protection; Database Producers; *Databases; Evaluation Methods; Foreign Countries; *Information Dissemination; Information Storage; Measurement Techniques; *Merchandise Information; Online Vendors; *Quality Control; Standards; Users (Information)
IDENTIFIERS Barriers to Implementation; Wales

ABSTRACT

The Centre for Information Quality Management (CIQM) was set up by the Library Association and UK (United Kingdom) Online User Group to act as a clearinghouse to which database users may report problems relating to the quality of any aspect of a database being used. CIQM acts as an intermediary between the user and information provider in obtaining solutions and collects statistics on database quality issues which they provide to the information industry. CIQM has proposed "Data Labelling" as a means by which users can be made aware of database capabilities and limitations. Database Labels are short specifications that include a qualitative assessment of a database's performance. Labels would be created by the information provider and include a complete statement of subject coverage, the total number of records, detailed geographic, language and time coverage, and simple statements of policy on points such as indexing and inclusion. Labels would have a uniform appearance in order to distinguish them from other documentation, and would be generated regularly, ideally with each product update. If Labels were accredited by an impartial agency, their value would be significantly enhanced, and Labels would then serve as a guarantee of product quality. Ways to implement labeling, implications, and barriers are discussed. (SWC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 405 880

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Database Quality: Label or Liable

by C.J. Armstrong

IR056309

BEST COPY AVAILABLE

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
C.J. Armstrong
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Database Quality: Label or Liable

C. J. Armstrong
The Centre for Information Quality Management (CIQM)

Database Quality

Amongst a lot of recent talk, articles and papers about quality in the information industry, an initiative by two professional organisations has already gone a long way in helping users cope with quality issues and, at the same time, has begun looking for a means of providing some security for future database users. The Centre for Information Quality Management (CIQM) was set up by the Library Association and the UK Online User Group to act as a clearing house to which database users may report problems relating to the quality of any aspect of a database being used (search software, data, indexing, documentation, training). CIQM undertakes to forward the problem to the appropriate body (information provider, online host, CD-ROM publisher) and route the response back to the user. This activity enables the collection of statistics on database quality issues which are fed back into the information industry. The service is free to users.

The overall objective of the Centre is to improve the quality of databases (online, CD-ROM, diskette, tape) and, in so doing, work towards developing a set of metrics by which database quality can be measured. Funding from the British Library Research & Development Department has enabled the Centre to begin work in this area and the remainder of this paper explores one possible methodology which offers users guaranteed performance levels for databases.

Currently, users have no knowledge of the formal specification for a database they are using - in effect, they are paying for an unknown quantity. Added to this, publicity material frequently generates unrealistic expectations that are not met when searching at the terminal. More reasonable expectations - for example, that authority files are used in the generation of primary index fields - are not always met either. No database so far evaluated at CIQM has standardised publisher names; this means that users frequently need to search for both 'John Wiley' and 'Wiley, John', for example. In one database the place of publication index contained over 40 variations on London including mis-spellings, concatenated MARC fields, and comments - 'Lond', 'Londin', 'LondonbRoutledge' (the 'b' is the remains of the '$b' sub-field marker), 'London sic', etc.

Many of the quality issues reported to CIQM reflect this gap in expectations and there seems to be a clear need - as a part of any drive to improve database quality - to develop a means by which users are made aware of database capabilities. The means being investigated at CIQM is Database Labelling.

Database Labelling

Database Labelling was first suggested by Péter Jacsó in a guest editorial in *Database* as analogous to food and drug labelling (Jacsó, 1993). Database Labels are short specifications which include some qualitative assessment of a database's performance. They offer potential users a means whereby they can determine exactly what is in a database and whether they want to use it: the extent to which they can 'trust' it.

The brief current description is supplied or created by the database owner/information provider and summarises the more complete and lengthy documentation in a way that users would find both easy to understand and accessible: a 'Contents List' supplied in a standard, recognisable format. One possible example is given in Jacsó's article.

On the one hand, the Label would supply a database specification including a complete statement of subject coverage (perhaps in the form of a topic list), the total number of records, detailed geographic, language and time coverage, and simple statements of policy on such points as indexing and inclusion. On the other hand, some measure of these might be given by noting the numbers of records against years, countries and languages, the average numbers of descriptors per record, and percentages for information points such as records with abstracts.

Factual information, such as number of records, geographical coverage, subject description or available fields, is supplemented by qualitative information which qualifies it: thus, geographical coverage could include the percentages of records for each

country and the list of available fields could include the number (or percentage) of records with actual data in each of the field types.

The Label would immediately show exactly what a database could do for users, leaving them with no unreasonable expectations. The Label would become a quality assurance statement demonstrating to what extent the database could be relied upon or 'trusted'. The factual information would give unambiguous parameters for coverage and use while the qualitative metrics would demonstrate how well the database functioned in these areas.

The Label removes the possibility of unsubstantiated marketing claims such as, 'The database has 26 access points' (indexes to be used in searching) which can no longer disguise the fact that - as has often been found - many of the 26 indexes do not contain data from every record. If an indexed field has only been filled for 80% of the records this will show on the Label.

Databases appear on different online hosts or CD-ROMs and may have a quite different appearance in each version. Different fields may be made available (with or without abstracts, for example), the indexing is generated by the vendor, print formats will almost certainly vary and software-related aspects which affect access and ease of use are certain to differ. For these reasons, Labels for each manifestation of the database will have to be generated - probably as a joint effort which involves both the information provider and the vendor/publisher.

Labels must have a uniform appearance in order to distinguish them from other documentation and a standard layout will make their use by users and prospective users simpler - comparisons can be made more easily. Some form of branding on the Label, for example by incorporating the CIQM logo, might be appropriate as it would mean that users could readily identify an independent 'Label' from other sales or marketing literature from the producer.

Effectively, the Label would become a database-specific standard. However, in using the term, 'standard', care has to be taken to distinguish between a Standard as defined by BSI or ISO procedures and the idea of an entirely local standard (or level of quality) which is specific to a given product. The information provider would specify database parameters as they pertain to a database at the point in time that the Label is first generated and then seek to adhere to or better that performance.

To be effective, the Label should be generated regularly - ideally to coincide with the normal vendor update cycle - and should be circulated with publicity material and made available on exhibition stands. It must also be made available to prospective users - published - in some form.

Even as described so far, a Database Label would perform a useful function, demonstrating to users the exact performance level of any database and acting as a benchmark against which future performance can be tested by users and producers alike. If Labels were accredited by an impartial agency, their value would be significantly enhanced. Labelled databases would, in effect, have a guarantee of quality. The Label would be seen by the user as an independent assessment of the database offering them a security hitherto unavailable.

The Accreditation Body

Accreditation by means of the Labels offers users a guarantee of quality and producers a 'kite mark' to flag their database as trustworthy. In turn, accreditation implies the existence of a neutral body which would be responsible for the mechanism of Label provision, verification and publication.

One of the most apparent problems with Labelling is the amount of additional work thrust on information providers and vendors. Labels become far more viable in terms of the workload if the central body (perhaps CIQM in association with the Library Association) produces a form to be filled in by producers.

As has been suggested, all Labels should look identical to the user. Consistency of Labelling is desirable but different services and different types of data are designed to meet different needs. The central body - liaising with database producers, hosts and publishers - will first need to take responsibility for developing a format for the Label and for producing guidelines as to what information should be put against the headings. It would be, in essence, a blank form which producers then fill in. It is not possible to define a single, standard dataset that can be applied to all databases; each database is different (bibliographic, image or text, for example) so it is not practicable to use one form of Label for all. A more pragmatic approach using a standard core of headings with options for the producer's own information or different Labels for different type of databases, might be more practicable.

In addition to specifying headings on the form for what should be included on the Label - for example, the number of records, coverage, fields, indexing, or publication years, definitions or 'scope notes' rendering the form easy to complete will be required. It

is essential that the task of producing the data for the Labels is simplified and automated as far as possible so that the information providers and vendors are able to supply the information regularly without detriment to their database production schedules. It may be most convenient for forms to be generated and returned electronically.

Once a database producer and database publisher have filled in the "form", it would be submitted to CIQM for audit and checking. When they have been approved these Labels could then be published and/or distributed to users by CIQM or some other publishing body. Simplicity is vital if the Label is to be of real help to users of a database. After the Label has been issued, the database will have to be periodically checked against the Label and the Label updated to ensure that it continues to accurately reflect the content and nature of the database. Periodically, new Labels will be published.

The mechanism for publishing the Labels has yet to be decided but, apart from making copies available to the information owner and the vendor to be distributed with documentation and publicity material, a means has to be identified which will make the Label readily available to any existing or potential user. The Internet may offer the most appropriate channel. Additionally, it is hoped that publishers of independent database directories might flag accredited databases in some way.

Will Labels Work?

In setting out this methodology for database quality assurance and in describing the possible advantages, it is important not to overlook the cost element - which would fall largely to the information provider - and other issues of use.

Labels must provide an accurate picture of a database as it exists when the Label is created or updated. Many of the major and most-used databases have been available electronically for 20 or more years and in this time have changed considerably. New fields may have been added (for example, an abstract) or fields may have been divided up to provide better access (Source field divided into Journal, Publication Year, Volume, Issue, etc, for example); thesaural control may have been introduced at some point; and coverage will almost certainly have improved. To give 'scores' representing the entirety of the database would give a false or a skewed impression of current production. It is not sufficient, for example, to show that 80% of the total content is from the United States when the average update since 1995 is 50% from USA, 20% from the UK, with the remaining 30% from continental Europe. One solution may be to show the dates of change: the date that fields came into existence and their rating for use in records from that date only, for example.

Unlike some publicity material and database factsheets, the Labels will need to be completely re-produced or updated several times each year; this clearly has considerable overheads in terms of both time and costs. Updating such Labels for all of a producer's databases in all their various forms would be a major task. It will certainly be necessary to date the Labels clearly on the front in order that users can see clearly that they are using a relevant and current version.

The volume of data to be condensed into a relatively small amount of space - no more than four A4 pages - is also problematic. It may be possible to balance the short, summary Labels with documentation made available electronically - possibly via the Internet - with links from individual databases. This is already happening to some extent; for example SilverPlatter has made available a free database of software parameters, hardware specifications and database details on their homepage.

A further consideration is the increasing use of databases distributed over local area networks (for example, in universities); how are the many users (many of them vulnerable end users) to be presented with the Labels. Users in any situation cannot be *made* to read the Label but it will be necessary to make users aware of the possibilities for quality control that are open to them. Local training and publicity supplied by library staff can back up efforts made by the information providers but the most useful tool may well be a logon message asking, 'Have You Read The Label?'

The Future

Database Labelling offers considerable benefits to users but will require a not inconsiderable infrastructure to function. Is it all possible? There is a huge backlog of databases to be 'Labelled' and a feasibility study will be necessary to assess the scale of the project. The consensus of opinion at a meeting of information providers earlier this year was that, at the very least, some preliminary research should be undertaken.

Future work at the Centre for Information Quality Management will aim to:

- raise the level of awareness of its aims and activities amongst users and the information industry

- gather more information from users across Europe on what they consider to be important quality issues as well as on the efficacy of Database Labelling
- develop a design for the Labels and the input form (complete with scope notes), and will
- set up feasibility and pilot studies to look at the mechanisms for the various stages of Labelling and the costs involved for both an accreditation body and the database industry.

It may be that a part of the infrastructure ultimately involves legal requirements to Label databases or it may be that Labelling progresses naturally due to peer and user pressures. One thing does seem clear: if the scheme goes ahead, the unaccredited databases will tend to lose marketshare to those that are accredited while the Labelled databases will be less liable to complaints from users - the Labels will ensure that users have no misconceptions about database scope and capabilities at the same time that the Label's benchmarking role gradually drives quality up.

Reference

Jacsó, P. A. (1993) 'Proposal for database "nutrition and ingredient" labeling'. *Database* 16(1) 7-9

6